DOCUMENT RESUME

ED 272 515 TM 860 268

AUTHOR Hsu, Tse-chi; Sadock, Shula F.
TITLE Computer-Assisted Test Construction: The State of the Art.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO ERIC-TME-R-88
PUB DATE Nov 85
CONTRACT 400-83-0015
NOTE 89p.
AVAILABLE FROM ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, Princeton, NJ 08541 ($7.50).
PUB TYPE Information Analyses - ERIC Information Analysis Products (071)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Adaptive Testing; *Computer Assisted Testing; *Computers; Educational Research; Educational Testing; Item Banks; *Measurement Objectives; *Measurement Techniques; State of the Art Reviews; *Test Construction; *Test Items

ABSTRACT

This report provides an overview of the current applications of computer technology to construct test items and/or to formulate tests according to sound measurement principles. The test items may be computer-generated from strategies programmed by test constructors, or pre-constructed by item writers and stored in computer memory. The tests formulated may be administered interactively by the computer or as paper and pencil tests. Studies dealing with computer applications in item construction, item banking, test design, and test administration (both adaptive and nonadaptive) are grouped for review in four sections: (1) theoretical and philosophical propositions; (2) applications and implementations; (3) evaluation and research; and (4) prospects for the future and implications for educational testing. It is concluded that while there have been many attempts to utilize computers for test construction, actual successful, large scale applications are relatively few. Most of these simply use computers to replace pencil and paper tests or human labor. With the exception of adaptive testing, there is little documentation to show that the quality of assessment processes is improved by computer utilization. However, with continuing rapid technological developments to overcome current computer limitations, and with attention to measurement quality, the future of computer-assisted test construction should be very bright. (BS)

COMPUTER-ASSISTED TEST CONSTRUCTION:

A State of the Art

TME REPORT 88

by

Tse-chi Hsu
University of Pittsburgh

Shula F. Sadock
Pittsburgh Board of Education

ERIC®

TESTS MEASUREMENT EVALUATION

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541-0001

COMPUTER-ASSISTED TEST CONSTRUCTION: THE STATE OF THE ART

by

Tse-chi Hsu
University of Pittsburgh

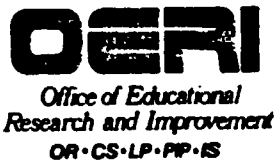
Shula F. Sadock
Pittsburgh Board of Education

November, 1985

ERIC Clearinghouse on Tests, Measurement, and Evaluation
Educational Testing Service, Princeton, New Jersey 08541-001

The material in this publication was prepared pursuant to a contract with the Office of Educational Research and Improvement, U.S. Department of Education. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Prior to publication, the manuscript was submitted to qualified professionals for critical review and determination of professional competence. This publication has met such standards. Points of view or opinions, however, do not necessarily represent the official view or opinions of either these reviewers or the Office of Educational Research and Improvement.

ERIC Clearinghouse on Tests, Measurement, and Evaluation
Educational Testing Service
Princeton, NJ 08541



This publication was prepared with funding from the Office of Educational Research and Improvement, U.S. Department of Education under contract No. NIE-400-83-0015. The opinions expressed in this report do not necessarily reflect the positions or policies of OERI or the Department of Education.

Table of Contents

I. INTRODUCTION

Educators have attempted to apply computer technology to testing since the emergence of computers. The earliest and most successful aplications are probably in test scoring, test reporting, and item analysis (Baker, 1971). Although many attempts have been made to apply computers to other aspects of testing, the degrees of success vary. In this paper, we will not attempt to provide a complete account of computer testing history. Rather, we will try to give a summary of the state of the art of computer-assisted test construction. We hope that the summary will be useful to the developers, researchers, and implementers of computer-assisted test construction systems.

Before we proceed to the main theme of the paper, however, we must describe our concept of computer-assisted test construction, because the term has been used to represent different activities by different people. Our concept of computer-assisted test construction includes any activity which involves the utilization of computer technology to construct items and/or to select a set of pre-constructed items to form a test. This concept emphasizes the application of computers to assist in the selection of items based on sound measurement principles. The items may be pre-constructed by item writers and stored in the memory of computers or generated by the computer from strategies programmed by test constructors. The test formulated through this process may be administered to pupils by the computer interactively or printed on paper and

administered as a paper and pencil test. Using this concept, we will review only studies dealing with item construction, item banking, test design, and test administration, either adaptive or nonadaptive.

Item construction concerns the utilization of computers in constructing or generating test items. Item banking deals with the systematic storage and subsequent retrieval and/or modification of previously constructed items. Consequently, our emphasis here is on item banking systems. If only item attributes, such as identification numbers and statistics, instead of items per se, are stored in the bank for the purpose of selecting items for a test, the process is considered in the category of test design. Item banks that involved no item classification and/or item selection strategies will be excluded. Test Administration includes applications utilizing computers to identify items from a larger item pool and administering the items to the students. The emphasis is on whether the computer is utilized to improve the quality of test administration. Using this criterion, we may include the majority of adaptive testing systems. Computer-assisted nonadaptive testing systems are included only if they appear to offer some advantages over the traditional paper and pencil, group administration approach. Thus, the administration of standardized tests on computers will be included. Tests administered as part of computer-assisted interaction (CAI) lessons will not be discussed because they cannot be considered independently from the CAI strategies, which are not the primary concerns of this article.

In addition to classifying studies according to item construction, item banking, test design, and test administration categories, we have further grouped them into four sections in the review. Studies dealing with theoretical and/or philosophical propositions will be grouped into Section II. Some of the ideas described in this section may have been incorporated into practices already. Others may still be in the stage of experimental trials. The objective of this section is to present the researcher's concepts of how computers should be utilized.

Section III contains applications implemented on both mainframes and microcomputers. Many of the applications appearing in the 1970s were designed for mainframes, while the 1980s are characterized by applications designed for microcomputers. Since the applications before microcomputers have a long history, a great deal of literature can be included in this section. Emphasis will be placed on applications implemented after approximately 1973. Studies published before that will be included only if they have implications to later developments. Readers interested in earlier developments may wish to consult Lippey (1974), Byrne (1976), and the 1973 special issue of Educational Technology on computer-assisted test construction. Most of the applications which emerged during the last few years involve microcomputers. Since microcomputers are so popular these days, it is important to have a good assessment of the status of these applications.

Section IV consists of evaluation and research issues associated with test construction applications. These issues may be related to either mainframes or microcomputers. Studies dealing with empirical investigations of theoretical issues or with evaluations of various applications may be included also.

Prospects for the future will be discussed in Section V. It will include a survey of prospects offered by researchers and our observations. Implications for the future of educational testing will be described also.

II. THEORETICAL AND PHILOSOPHICAL PROPOSITIONS

Ideas about how computers should be used in test construction are the seeds for innovations. In this section, we are going to summarize some of the ideas appearing in recent literature according to the four categories posted previously. Readers seeking additional ideas may also consult Baker (in preparation); Hambleton (1984); Hambleton, Anderson, and Murray (1983); Oosterhof & Salisbury (1985); Roid (1984b); and Sampson (1983).

Item Construction

Using computers to construct items is not a new concept. Anastasio and his associates attempted to use computers to construct sentence completion and spelling items in the late 1960s (Anastasio, Marcotte, & Fremer 1969; Fremer & Anastasio, 1969). These works, however, were never really adopted by test constructors. Several researchers attempted to generate items in the early 1970s (Ferguson & Hsu, 1971; Feldker, 1973; Vickers, 1973). But the strategies of item generation are different. One of the most common approaches is to generate items based on item forms which represent a specific domain of contents.

A prerequisite for using computers to construct items is to develop algorithms for item construction. These algorithms must be based on various item writing techniques. Although the

interest in item writing techniques is not new, recent interest in this topic focuses on how these techniques may be computerized. For example, Hsu (1975) discussed four achievement test construction approaches: Guttman's facet design, Hively's item form analysis, Scandura's algorithmic analysis, and Bormuth's operational approach (or linguistic transformation). These and other techniques are illustrated and discussed in detail by Millman (1974), Roid and Haladyna (1982), and Roid (1984a). Attempts to computerize some of these techniques have been made. For example, item form analysis has been implemented in Ferguson and Hsu (1971), Hsu and Carlson (1973), and Millman and Outlaw (1978). The facet design has been tried by Berk (1978). The linguistic transformation approach has been utilized by Finn (1975) and Roid and Finn (1978). Some of these applications will be described in more detail later.

Item Banking

An item bank is a collection of items that has been organized and classified in terms of the content and/or the statistical characteristics of the items. Most banked items are objective items such as multiple-choice and true-false. In this section we are not concerned with describing existing item banks. Rather, our emphasis is on computerized item banking systems. The purpose of an item banking system is to catalogue, modify and maintain a set of items. Before developing an item

bank, one may want to make sure whether an item bank is really needed. Millman and Arter (1984) pointed out that for an item bank to be valuable, at least one of the following conditions must be met:

1. Tests constructed according to local specifications are needed and not yet available;
2. Frequent testing is required;
3. Multiple forms of a test are needed;
4. Individually tailored tests are desired;
5. Multiple users and/or contributors are willing to cooperate;

   and/or

6. An item banking system is available.

An item bank to be useful for test construction, however, is not easy to design (Hiscox, 1984a). Several factors must be considered. First of all, items in the bank must be classified meaningfully and systematically. Item classification systems should not be exclusively governed by concerns for quick and efficient retrieval of items. Rather, proper classification should aid in improving the validity of the test to be constructed. The item classification system should be dependent upon the purpose of the item bank. We cannot expect one classification scheme to be used for all purposes.

The second criterion which should be considered is whether the bank is easy to maintain. This may include procedures for creating, storing, retrieving, and modifying items. Some word processing capability is desirable. But this capability should not occupy space needed for manipulating the item bank.

Item evaluative data are needed for selecting good quality items. A good bank should be used to maintain and to update these data for the users. Since there are many varieties of item data, the bank should be flexible in terms of the kinds of statistics needed by different users. It will be most desirable if the users have the option to choose the kinds of item statistics to meet their needs.

Another factor which should be considered is whether the procedure for assembling items is adequate. A commonly used procedure is the selection of items one at a time by the users. One advantage of this approach is that the user has a chance to evaluate each item carefully and then decide whether the item should be used. This approach, however, is very inefficient especially when the bank is relatively large. If possible, items meeting criteria specified by the users should be selected first before examining items one at a time. Random selection of items without users' review is not desirable.

The final criterion which should be considered is the flexibility of the item bank usage. In assembling the tests to be printed, who will decide the order of the items to be printed? Is it possible to use the same bank for on-line testing? Is it possible to print items directly on stencils or dittos so that many copies of the tests can be made easily?

Although the technical quality of programming is desirable, the technical quality of testing should not be sacrificed. Item banking is not simply a means of storing items, but should

assist users in selecting high quality items for a specific purpose. Therefore, one or two seconds delay in retrieving an item is probably not as important as the question of whether this is the most desirable item for this purpose. Does the bank incorporate enough measurement principles so that it can provide sufficient clues to the users about the quality of the item? This should be the primary consideration in designing an item bank.

In addition to the criteria mentioned above, other issues regarding item banking can be found in Estes and Arter (1984) and Millman and Arter (1984). Besides describing the advantages and the disadvantages of an item bank, the last reference also provides an extensive list of questions to be answered when designing an item bank and determining the type of item information that may be stored in the bank. Readers interested in designing an item bank based on the Rasch model may consult Wright and Bell (1984).

Test Design

The category of test design considers the test as a whole. The primary concern of a test, of course, is its quality. Two major indicators of quality are validity and reliability. How computers may be used to design a test and judge its quality is the central theme of this section. In order to design a test and judge its quality, item and test statistics should be computed and evaluated. Since using computers to generate

item/test statistics is not included here, we focus only on the evaluation of item/test statistics for the purpose of item selection.

In planning for a classroom test, Nitko (1983) suggested the following major steps:

1. Define the purpose for testing at this time.
2. Specify the performance and processes to be observed and tested.
3. Select the type of test items or the methods to be used to observe and to test the performance.
4. Develop the initial drafts of the test exercises.
5. Are the items of satisfactory quality? If not, revise or reconstruct the items.
6. Do the items match the stated performance to be assessed? If not, revise or reconstruct the items.
7. Conduct a preliminary tryout of items, if possible.
8. Do the items appear to be functioning as intended? If not, revise or reconstruct the items.
9. Develop the final version of the test.
10. Administer the test and analyze the results.
11. Does the test appear to be functioning well? If not, revise or reconstruct the items.
12. Use the test for decision-making.

We cannot expect the computer to assist us in all 12 steps. But an innovative researcher may be able to find some ways to utilize the computer for certain functions in each

step. One possible exception is probably Step 1, defining the purpose of testing. Steps that are most relevant to test design are 3, 4, 6, 7, and 9. Step 5 is also a part of test design. In this article, however, it is classified in the category of item banking, which maintains actual items.

At Steps 3 and 4, computers may be utilized to assemble items according to content and/or item types. At Steps 6, 7, and 9, the computer may be used to store, compute, and display the item/test statistics needed for judging the quality of items/test. For example, based on item difficulty and discrimination indices obtained from previous testings, an estimate of the reliability coefficient of the new test can be made. This procedure has been implemented in the Pittsburgh Educational Testing Aids system (Nitko & Hsu, 1984a).

If estimators of item parameters based on item response theory (IRT) are available, the potential of estimating the quality of a new test is even greater. Since IRT offers a means by which item and test characteristics can be independent of the performance of some tryout group, "it becomes possible to describe in precise terms the characteristics of the test before the test is administered. This capacity allows one to construct a test that is highly efficient in accomplishing the purpose of the test" (Warm, 1978, p. 17).

Many different IRT models have been proposed for designing tests including the one-, two-, and three-parameter models. Comparative studies (e.g., Koch & Reckase, 1978, 1979; Urry, 1970, 1977) have investigated the utility of employing the one-

and three-parameter (1PL and 3PL) models. Although the 3PL model yields higher reliability, it is prone to nonconvergence of ability estimates. Nonconvergence is rarely a problem with the 1PL model and McKinley and Reckase (1984) recommend this model when small item pools are used.

Numerous calibration procedures exist for obtaining item parameters and ability estimates (e.g., BICAL by Wright, Mead, & Bell, 1980; BILOG by Bock & Aitkin, 1981; and LOGIST by Wingersky, Barton, & Lord, 1982). The mathematical complexity of the models, however, necessitates the use of a mainframe computer to obtain these estimates. Many attempts have been made to computerize the test design applications of IRT. While some systems have incorporated the parameter estimation procedures, other approaches rely on precalibrated item data (e.g., Holmes, 1983; Sadock, 1984). Before presenting some computerized applications of IRT, we will examine the theoretical basis of some of these test design systems. Following the test development applications of IRT, Lord suggests the following approach for designing a mastery test:

1. Obtain a pool of items for measuring the skill of interest.
2. Calibrate the items on a convenient sample.
3. Considering the entire item pool as a single test, calculate the test characteristic curve.
4. Define mastery in terms of true score.
5. Find the o-cutoff equivalent using the mastery score and the test characteristic curve.

6. Evaluate the item information at the o-cutoff equivalent.
7. Decide what length confidence interval for o will be adequate at the cutoff equivalent. Using this information, determine the required test information at the cutoff.
8. Select items with the most information at the cutoff and continue selecting until the sum of the item information at the cutoff equals the required test information.
9. Compute scoring weights for each item selected.
10. Compute the weighted sum of item scores for each examinee.
11. Compute the cutoff score.
12. Administer the test and select examinees with scores greater than the cutoff score. (Lord, 1980, pp. 174-175)

In computerizing the above procedures, several modifications have been made including provisions for specifying several cutoff scores (i.e., the design of classification tests) and the selection of items based on a maximum information range for each item. For example, the IRT-based test design system developed by Sadock (1984) allows users to specify up to four cutoff scores. In addition, under certain conditions, items are selected if their point of maximum information falls within a fixed range on the ability scale. A more precise indicator of

maximum item information has been proposed by Reckase & McKinley (1984). They suggest selecting items based on a variable length item effectiveness range. In addition, a new item difficulty parameter is defined as the midpoint of the effectiveness range.

Theoretical investigations by Samejima (1977) and Thissen (1976) have resulted in several new IRT models which incorporate information available from incorrect responses in estimating ability levels. Using response characteristic curves and response information curves, Woods (1983) describes a computer-aided item development procedure based on Samejima's models. IRT applications for both item construction and test design are used in this procedure.

The theoretical IRT literature is quite extensive. Over the past five years, much of the IRT literature has been centered around test development applications. The mathematical complexities of the IRT models, in many cases, require the use of a computer in applying the models to practical testing situations. Although many packages exist for calibrating items on mainframe computers, few calibration procedures exist for microcomputers. Baker (in preparation) has identified one microcomputer calibration system. MICROSCALE (Mediax, 1984) is a microcomputer version of the BICAL program which is based on the Rasch model.

The memory size and speed of micros significantly contribute to the current lack of software. However, because of the increasing use of micros in designing tests, as well as increased availability of add-on memory, we see a need for

procedures for approximating item and test parameters which may be implemented on microcomputers. Microcomputer-assisted test design systems which currently employ IRT principles include calibrated item data in the item bank thereby eliminating the need to calibrate the items with the micro. However, this approach requires the use of both mainframe and micro in developing such a system.

Test Administration

Administration of tests by the computer is justifiable only if it can improve the quality of testing. For example, the quality of testing can be improved by using the computer: (a) to provide immediate feedback, (b) to select the next item based on the response, (c) to store and analyze test results, and (d) to increase test security. But there are also many difficulties in administering tests using the computer such as: (a) the need for one computer (or terminal) for each student, (b) difficulty in tracking omitted items for review once all items have been attempted, (c) limited space in one screen, (d) limited memory storage, (e) slow speed printers for printing test results, and (f) difficulty in overcoming "computer phobia" by some examinees.

Two issues related to test administration will be discussed in this section. The first issue concerns adaptive testing, which became feasible because of the availability of the computer. The second issue addresses the impact of technology of test administration.

The major advantage of adaptive testing is the reduction of test administration time without the sacrifice of measurement precision (W. C. Ward, 1984). Adaptive testing procedures consist of three components: an item selection routine, an ability estimation technique, and a stopping rule.

Stocking and Swanson (1979) outline the typical adaptive testing algorithm as follows:

1. Obtain an initial estimate of the examinee's ability level.
2. Use this estimate to select an appropriate item from the item pool.
3. Administer and score the item. Use this information to revise the estimate of trait level.
4. If the estimate is satisfactory, stop. Otherwise, further refine the estimate by returning to Step 2.

Several procedures exist for selecting appropriate items (e.g., Samejima, 1977, Lord, 1971, and Wald, 1947). Typically, an incorrect response is followed by the administration of an easier item, and a correct response is followed by a more difficult item. The precise characteristics of the item (i.e., item difficulty and information) are dependent upon the selection algorithm. For example, selection rules may include, but are not limited to, the following procedures:

(a) For Bayesian updating, items with the highest discriminating power are selected since these items reduce the posterior variance.

(b) Items with maximal information are selected when maximum likelihood ability estimation is used. (Green, Bock, Humphreys, Linn, & Reckase, 1982)

Ability estimation procedures are numerous including the confidence interval approach, point estimation based on regression, maximum likelihood estimation, and Bayesian estimation approach (Weiss, 1974).

Although many stopping rules have been used, the three most frequently implemented rules include the following:

1. Stop when a fixed number of items has been administered,
2. Stop when all items with maximum information at the current ability estimate have been administered, and
3. Stop when a stable ability estimate has been obtained.

Because of complicated ability estimation procedures used in adaptive testing, most computer applications of adaptive testing have been implmented on mainframe computers.

Implementation of adaptive testing on microcomputers should be possible with the new development in hardware and software. Baker (1984) pointed out several technological trends, which should have implications on test administration: (a) 32 bit internal registers and 16 bit addresses and data bases, (b) new optical storage devices, (c) video disks, (d) device to scan graphic material and software to create graphic material, (e) voice input and output devices, (f) computer networking, and (g) new software development. These trends may not only enhance the capability of test administration, they may also help to revolutionize the types of test items.

In designing both adaptive and noadaptive tests to be administered by computers, several guidelines are available (Brightman, Freeman & Lewis, 1984; Mizokawa & Hamlin, 1984; Wedman & Stefanich, 1984). These guidelines argue how instructional strategies, psychometric theory, and technology may work together to administer tests effectively. In general, technology should be used to facilitate rather than to handicap students' responding processes. Therefore, the format, the rate, and task demands of item presentation should be designed in such a way that errors in using hardware to respond can be minimized.

III. APPLICATIONS AND IMPLEMENTATIONS

Item Construction

Many computer-assisted test construction packages which do not include actual items are equipped with item generators. There are two basic types of item generators: (a) those that store parts of an item (e.g., the stem and the options) and include rules for combining these parts to construct a whole item and (b) item generators based on item forms which consist of rules for generating items.

This first type of item generator was used in the Question Pool Management System (QPMS) (Denny, 1973). This system requires users to break up each item into three components: the item stem, seven possible correct answers, and seven possible distractors. For any item selected for inclusion in a test, the generator randomly selects one correct option and four distractors.

The more frequently used computerized method of constructing items relies on item generators which construct items from item forms (Hively, Patterson, & Page, 1968). The Individually Prescribed Instruction (IPI) mathematics programs (Hsu & Carlson, 1973) used this approach. Each unit was comprised of several objectives and each objective was divided into several item forms.

Millman and Outlaw (1978) have developed a system which enables users to construct test items using item programs.

These item programs are another type of item generator. The item programs must be written by the user using an expanded version of BASIC. Using a sample item as a guide, variations of the item are then produced by breaking up the item into logical segments. For each segment, sets of alternate words are constructed. For instance if an item is to test knowledge of characteristics of different plants, the item program may be written to select among five different plants and three different characteristics. With the item program, the user can specify when a segment is to be randomly generated from among the segments or if the generation of one segment is dependent upon the variations previously selected. Possible answers are also selected by the item program. Answer choices can also be selected on a random or conditional basis. This system has been discussed and illustrated in Millman (1980, 1982).

Instead of merely using computers to perform permutations and combinations of content preconstructed by the item writer, Millman and Westman (J. Millman, personal communication, 1985) are designing an improved system which incorporates some artifical intelligence capabilities. In using this system, item writers can be assisted by the system interactively in various ways: (a) examining prototype items measuring the desired processes, (b) offering prompts based on users' needs, and (c) accessing a number of available system libraries.

So far, we did not find any application of microcomputers in item construction in the literature. This is probably due to the limitation of storage space of the first generation of

microcomputers. With the development of the second generation of microcomputers and the progress in artifical intelligence, we may anticipate more applications in this area (Roid, 1984a). Although several articles discuss the possibility of automating item construction processes (e.g., Millman, 1980), more theoretical work in item writing techniques is still needed.

Item Banking

Item banking systems tend to fall into one of two categories: those equipped with and designed for a specific set of items and those that require users to enter and store their own items. To be included in this section, an item banking system must meet the requirements of flexibility stated earlier

MEDSIRCH, used by medical schools in Canada (Hazlett, 1973), is an item banking system which requires users to create and store their own items. MEDSIRCH allows a maximum of 57 variables or codes to describe each item. These include such things as area of specialty and subspecialty, degree of importance, difficulty level, history of use, and even type of audiovisual equipment necessary. The user creates the bank by preparing items on keypunch cards and submitting the cards to a series of programs which check, catalogue, and eventually store the items on tape. A similar system developed at Iowa State University (Menne, 1973) was implemented because many instructors were filling out their own items on index cards or IBM cards and using the text editor of the university's computer

to prepare tests from the item cards. With this system users have the option of establishing their own item banks or using one of six existing banks on the system.

TICAT (Tuskegee Institute Computer-Assisted Tester), an interactive computer-assisted testing system can be used to develop item banks, create and administer tests, and score and report test results (Howze, 1978). The system was written in Time-Share BASIC for a Hewlett-Packard 2000-ACCESS system. The item bank component can be used to develop item files containing a maximum of 128 true-false or multiple-choice items. An item file is the actual set of items for an examination. It is not the set of items from which the exam items will be selected. When items are entered in the file, the system prompts users for the item text, the correct answer, and a citation or textbook reference. Storage space is reserved for additional information regarding item usage. Item information may also be edited and updated. A COPY routine can be used to combine all or portions of item files together to form new files. Since item files are actually exams, this feature allows considerable flexibility in combining several sets of items covering different content areas. A LIST command, which allows users to print all or part of an item file, is also included. This command, however, does not produce the final printed copy of the test. This function is performed by a separate component of the system.

Both a manual and a computerized item bank can be used with CIBEDS, the Computerized Item Banking and Exam Development System (Vale, 1979). The manual card file was developed as a

first step in designing the bank. The bank consists of approximately 15,000 items used by the Minnesota State Department of Personnel for personnel classification tests. Items were classified using a nine-digit code based on the Dewey Decimal system. The computerized bank (written in FORTRAN for the Control Data Cyber 75 time-sharing system) is capable of storing items and item statistics, modifying items, selecting items based on item content and/or item statistics, and formatting items in a photocopy-ready form.

The computer-assisted test construction system developed by Stock, Esterson, and Schmid (1977), has integrated both item banking and test design capabilities. In creating the item bank, the user is required to define a two-way test specification table. Items are then referenced to this test blueprint. The system is capable of (a) adding items to the bank via batch processing, (b) generating tests by obtaining a stratifiedrandom sample of items from the user-selected cells of the test plan, and (c) editing the item bank. The item bank contains item analysis data including item difficulty and discrimination, the keyed response, and a reference to the table of specifications. Items are selected in an interactive mode, whereby users specify cells in the specifications table and the program responds by indicating the number of items in the bank classified for each selected cell. The system is also capable of generating parallel forms of the test. The system was designed on a Univac 1110 Exec 8 system and includes an item bank containing over 900 measurement and statistics items

classified by 63 content areas and three skill levels (e.g., facts, principles, and applications).

The science question bank developed for the Assessment of Performance Unit in the Department of Education and Science, England, is a comprehensive system designed specifically for monitoring science performance (Johnson & Maher, 1982). The information retrieval system for this bank was implemented on the AMDAHL V/7 computer at the University of Leeds using the CODIL programming language. The system consists of (a) a BREAKDOWN procedure, which can summarize items according to major- and sub-categories, (b) a BROWSING procedure, which allows a user to specify the characteristics of items to be reviewed, and (c) a TEST CONSTRUCTION procedure, which assists a user to select items to form a test. A random question may be rejected if it violates any of the conditions (range, frequency, inclusive/exclusive) prescribed by the user. The browsing capability of the system has been enhanced greatly with the addition of a thesaurus access option to the system (Johnson & Maher, 1984).

Many microcomputer item banking systems appeared during the last few years. In their review, Deck and Estes (1984) have identified at least 75 packages. Since microcomputers are accessible to most people and most instructors need some place to store items, no wonder so many packages were developed. Several reviews on item banking for microcomputers have also appeared during the last few years (Deck & Estes, 1984; Hambleton, 1984; Hsu & Nitko, 1984). We will not attempt to

present an extensive review of existing microcomputer-assisted item banking packages here. Rather, in the remainder of this section we focus on several microcomputer applications to illustrate what we consider to be desirable characteristics of a microcomputer item bank that can be used to construct tests.

ITEMBANK, designed by Bowers (1984) at the American College Testing Program, is a relatively large and complicated item bank that may be appropriate for state and local agencies. This bank was established, maintained, and updated using dBASE II. The entire test development process was divided into four stages: the Draft Test Stage, the Working Test State, the Final Test Stage, and the Item Analysis and Updating Stage. There are sub-menus for each stage. The Working Test stage and the final test stage are also relevant to test construction. In the Working Test Stage, reviewers' comments are incorporated and new item data are entered into the data-base. Items are revised, assembled, and sent to reviewers for further evaluation. In the Final Test Stage, items are further revised. A scoring sheet and a test summary sheet are printed.

This bank was implemented on an IBM PC with color monitor. It is a good example of using a computer to assist in the construction of tests, not just using the computer to store items. This system may not be appropriate for individual teachers, however. In addition to the requirement for dBASE II software, teachers without computer training may find the system too complicated. In fact, most teachers may not really need such a large system. dBASE II also has been used by DeGruijter

(1985) to develop an item banking system which can incorporate test analysis results into the item bank.

Bowers also produced a newer version of the system, entitled dBANK, using dBASE III (Bowers, 1985). He claimed that dBANK should be considered as a totally new system. A primary difference between dBANK and its predecessor ITEMBANK is that dBANK emphasizes functions rather than tasks. The four menus in dBANK contain: (a) data file maintenance functions, (b) draft test development functions, (c) report printing functions, and (d) data communication functions. Another change is the redesign of programs in order to take advantage of dbase III's increased speed. Both screen displays and data base files were redesigned to take advantage of the new features and capabilities of dBASE III. In conclusion, the developer emphasized that "dBANK ensures data integrity and makes the data management and reporting associated with test development far more accurate and efficient." (p. 12) (Other systems of similar capabilities: W. H. Ward, 1984; Hiscox, 1984b.)

MicroCAT (Assessment Systems Corporation, 1985) and the Pittsburgh Educational Testing Aids (PETA) System (Nitko & Hsu, 1984a) are comprehensive testing packages which contain rather extensive item banking components. A somewhat detailed description of these item banking components is presented below. MicroCAT (Assessment Systems Corporation, 1985) can assist users to develop, administer, score, and analyze computerized tests. Designed for the IBM series of microcomputers (PC, XT, and AT), the package consists of four

subsystems: development, examination, assessment, and management. The development subsystem includes five programs entitled Graphics Item Banker, Font Generation, Test Specification, Text Editing and Test Compilation. The first two programs can be used to enter, retrieve, and modify test items and instructions. A brief description of their features appears below. The last three programs can be used to assemble tests. Their characteristics are summarized in the Test Design section. The examination subsystem can be used to administer a test to a single examinee or a group of examinees. Examinees' responses may be cumulated for later analyses by the assessment subsystem. The assessment subsystem can be used to evaluate the performance of items/tests using both classical item/test analyses and item response theory. Since these features and the functions of the management subsystem are not included in our definition of a computer-assisted test construction system, they are not discussed in this paper.

With the item banking programs (i.e., Graphics Item Banker program and the Font Generation Program) this system is capable of storing up to 14 item characteristics such as display time, correct response, and estimates of item parameters. In addition, 22 special graphics commands are available for creating graphic items. These commands can be grouped into five categories: geometric primitives for drawing standard shapes, text commands, additional drawing commands, graphics segmenting commands, and utilities commands. These commands should be sufficient for most common uses. Items entered by this system

also can be organized according to content areas and grouped into separate directories. This system has incorporated up-to-date measurement principles. However, it appears that a sufficient knowledge of measurement is required for efficient and effective use of this system.

The item banking component included in the Pittsburgh Educational Testing Aids (PETA) system (Nitko & Hsu, 1984b), is specifically designed for individual teachers. Although this component and two other components (i.e., student data-base and item analysis) form the PETA system, the item banking component can be used independently to maintain test items and to construct classroom tests. This system was implemented for the Apple II plus and Apple IIe with 48K memory. For item banking, two disk drives and a printer are required.

There are 11 programs in the item banking component plus one option for terminating the execution. Since the main menu reflects the capability of the system, the options available are listed here: 1. CREATE/ENTER INTO AN ITEM BANK: A TEST ITEM AND ITEM DATA; 2. CREATE/ENTER INTO AN ITEM BANK: A TEST DIRECTION; 3. RETRIEVE FROM AN ITEM BANK: A TEST ITEM AND ITEM DATA; 4. RETRIEVE FROM AN ITEM BANK: A TEST DIRECTION; 5. REORGANIZE AND COPY: AN ITEM BANK TO SAVE SPACE; 6. TRANSFER ITEM STATISTICAL DATA: FROM A DATA-BASE TO AN ITEM BANK; 7. TERMINATING; 8 RETRIEVE FROM ITEM BANK: ITEMS MEETING YOUR CRITERIA; 9. RETRIEVE FROM A TEST FILE: A TEST ITEM FILE AND ITEM DATA; 10. RETRIEVE FROM A TEST FILE: A TEST DIRECTION; 11. ESTIMATING THE PROPERTIES OF: THE TEST ON A TEST FILE

DISK; AND 12. PRINT ITEMS/DIRECTIONS IN AN: ITEM BANK AND/OR TEST FILE (Nitko & Hsu, 1984a).

The item bank component consists of several special features. It does not utilize separate word processors for item creation and modification. However, several special commands are included to facilitate word processing. This bank can store five most commonly used item types: multiple-choice, true-false, matching, fill-in, and essay questions. The maximum length for an item is 1522 characters (including spaces), but the random access file record length is only 122 characters. This implies that a longer item may be placed in more than one record. For shorter items, however, no wasting of space is necessary.

Since items may be retrieved either by criteria or item identification numbers (representing contents), the user has full control of the items tc be selected. In retrieving items according to criteria, the user may modify the criteria in order to increase or decrease the number of items to be assembled in the test file disk. The user may examine or modify an item either when it is in the bank or after it has been selected and placed into a test file. Item data, either classicial item statistics or estimated item parameters used in item response theory, are also presented to the user simultaneously with the item. When a test is assembled, the user may request to estimate the quality of the test using item statistics obtained previously. When a test is ready to be printed, items may be ordered according to content. difficulty, or any order specified by the user.

Since this system was designed specifically for individual teachers and restricted to a minimum hardware requirements, it cannot include any graphics and symbols in the items. The response time is not the fastest, but it is fast enough for most common usage. The classification scheme for items in the bank is especially appropriate for classroom testing.

Since the application of item response theory in building item banks is becoming popular, a framework proposed by Wright and Bell (1984) is described here. This framework has been used to build item banks used at several school sites. It consists of three components. The first component is Bank Plan. Program Form in this component is used to decide what item will be included in a particular form. The second component is Test Administration, where tests are administered externally and responses are obtained. The final component is Bank Building. In this component, the program FORCAL calibrates items using the Rasch model. Then a series of fit analyses are carried out by the program SHIFT. Formulas needed for the analyses are also provided. An important contribution of this framework is the emphasis on the psychometric aspect of item banking. Calibrations of items are incorporated into the process of bank building.

Test Design

The selection of items for a test is dependent upon the user's content specifications and the item selection algorithm

of the individual computer-assisted test construction system. The typical system requires users to indicate the number of items desired for their test and specify item selection restrictions by identifying the type of items needed. Generally, items may be specified using any variable by which the items are classified in the bank. The item bank is searched and all items which meet the user's criteria are noted. Most programs then randomly select from among the items that satisfy the user's restrictions (e.g., Baker, 1973; Libaw, 1973; Toggenburger, 1973). The level of specificity of restrictions is obviously related to the level of specificity of item classifications in the item bank. Programs with very crude classification systems allow the user to specify actual item numbers only (Brown, 1973; Menne, 1973). On the other hand, if the item bank has an elaborate classification system, many more restrictions may be specified.

Both MEDSIRCH (Hazlett, 1973) and CTSS (Toggenburger, 1973) prioritize item selection criteria. An initial search of the item bank is performed and the number of items meeting all criteria is noted. If the number of items satisfying these criteria is less than the requestednumber of items on the test, one criterion is dropped and the bank is searched again. This process is continued until enough items are identified for selection. The MEDSIRCH and CTSS packages differ in that MEDSIRCH allows users to prioritize their item selection criteria whereas CTSS has established its own prioritization. Behavioral level criterion, which CTSS defines as knowledge or

application of knowledge, is dropped first, followed by item difficulty. Sivertson, Hansen, and Schoenenberger (1973) describe a unique test design system designed to "identify the continuing education needs of individual physicians" (p. 38). Their comprehensive item bank contains 2020 five-option multiple-choice items covering "all diseases a physician might encounter in his practice" (p. 39). Items were first classified using the International Classification of Diseases, Adapted (ICDA) codings. The authors then added both specialty codes, such as General Practitioner (GP), Internal Medicine (IM), Pediatrics (P), and General Surgery (GS), and three skill level codes as follows:

Level 1: a common clinical situation and "on the spot" decision

Level 2: a decision requiring commonly available diagnostic tests and procedures

Level 3: a problem or technique requiring specialized training or diagnostic tests to manipulate information (p. 39).

To select a subset of items for a test the physician (user) must indicate his/her area of speciality. From the items that match the physician's specialty, a random sample of items is drawn from each skill category.

The SOCRATES' computer-assisted test retrieval system is an extensive computer network consisting of 11 item banks and over 10,000 items which are available throughout the 19 campuses of the California State University and Colleges system (Seely &

Willis, 1976). Items can be selected by subject category, difficulty level, behavior level (classified as either knowledge or application), and/or keyword. Maximum test length is set at 150 items. The system is capable of modifying a test 99 times and producing up to 10 scrambled forms of a test. Since the system is available to both faculty and students, students may request practice tests. The unique feature of this system is its networking component. When a test is designed the printed copy of the test can be produced at the site of origin if a high-speed printer is available. In addition, tests can be requested by telephone, assembled at the central processing site in Los Angeles and delivered to the campus via a courier service. Any campus which has a direct link with the central processor can design and print tests at that site.

In summary, the item selection strategy used by most earlier test design systems employing mainframes is to select items meeting various user-specified restrictions. When the system includes an item bank, the degree of specificity of test characteristics is dependent upon the classification scheme of the item bank.

Instead of selecting items based on user-specified item characteristics, item response theory may be employed to design tests. One example is the IRT Test Design System (Sadock, 1984). The system can be used on an Apple II Plus or Apple IIe with a minimum of 48K memory. One disk drive and a printer are required.

This system was designed as a tool for selecting a set of items for a relatively short, yet efficient test, without requiring that users possess a complete working knowledge of test development theory. There are five components to this system: (a) the test content specification component, (b) the test use component, (c) the test construction component, (d) the test modification component, and (e) the technical information component. Only the test use and test construction components are described here.

Similar to many test design systems, the test content is specified by selecting individual or groups of objectives or content areas. Many computer-assisted test construction systems proceed at this point by selecting a random sample of items appropriate to the content domain. The IRT System, however, requires that users indicate how the test scores will be used. This is accomplished via the test use component. At present, the system can design three types of tests, each serving a different purpose. The three test types include: (a) tests designed to group students (typically referred to as classification tests), (b) tests designed to rank students, and (c) tests designed to assess individual student mastery (typically referred to as objective mastery tests). The test use component provides users with nontechnical descriptions of these different test uses.

The item selection strategy for all three types of tests is based on the three-parameter logistic model. (Calibrated item data are included in the bank that accompanies this system.) When a grouping test is requested, the number of groups to be

formed as well as the percent cutoff-scores must be specified. Items are then selected if their point of maximum information falls within an acceptable range of the theta-equivalent of any cutscore.

When a test designed to rank students is requested, an estimate of the class ability level must be specified. This is necessary to insure that the items selected provide adequate information for ranking the entire group. By specifying class ability level, the user is actually specifying acceptable values of the maximum information levels of the items.

If a test designed to assess individual student mastery is requested, items of varying degrees of difficulty across all ability levels, but which explicitly represent the content characteristics of the domain, are selected. Note that there is no mastery cut score associated with this type of test. The purpose here is to determine the proportion of the content domain which each student has mastered.

At present, some components of this system are specific to the accompanying item bank. The second version of the system, however, will contain utility programs for adapting this system to any three parameter IRT calibrated item bank.

The Rasch model has also been applied to test design systems. The item bank and score conversion program described by Haksar (1983) were designed based on the Rasch criteria of an efficient test. From previous test results, the class score distribution (including the mean and standard deviation), can be approximated. According to the Rasch model, the test mean

should equal the average item difficulty, and the distance on the ability scale between the easiest and the most difficult item should be four times the standard deviation. Using these criteria as well as estimates of the score distribution, the user selects items from the item bank.

The item bank contains item difficulty measures in addition to content codes. At present the item bank is not in computer form. Rather, item selection is accomplished manually by inspecting either an item catalogue or an indexed set of item cards. The item cards are arranged in order of difficulty within content areas.

In order to translate a raw test score into a scaled score, a score conversion program has been written for an Apple. This program places the raw score onto the same ability/difficulty scale that was used in defining item difficulty.

One important feature of MicroCAT (Assessment Systems Corporation, 1985) is the test specification procedure. Six predefined templates are provided. The templates are incomplete test blueprints which enable the user to specify the items and requirements of the test. The templates permit the users to assemble a fixed-length conventional test, a variable length conventional test, or a variable-length adaptive testing using either Bayesian, maximum likelihood, or stratified adaptive decision strategy. Normally these templates require the user to identify items to be included and criteria required for the selected testing procedure. If none of the predefined templates is appropriate, users may create their own templates by using

the Minnesota Computerized Adaptive Testing Language. These new templates, however, must be compiled before they are used to design tests.

Test Administration

Adaptive testing. Many adaptive testing systems have been developed during the last decade (Clark, 1976; Weiss, 1978, 1980, 1983). Most of the systems were designed for research purposes rather than for actual implementation. Also, most of the systems concern aptitude measurement. Only recently has attention shifted to achievement testing (Bejar, Weiss, & Kingsbury, 1977; Brown & Weiss, 1977; Weiss & Kingsbury, 1984). Two examples of adaptive aptitude testing are discussed below. Adaptive achievement testing is discussed in more depth in the next section on Evaluation and Research.

Unlike many adaptive testing systems, TAILOR (Cudeck, Cliff, & Kehoe, 1977; McCormick & Cliff, 1977) does not require extensive pretesting of items. Rather, using the tailored testing approach by Cliff (1975), TAILOR estimates both item and person characteristics simultaneously. There are two versions of TAILOR. TAILOR-APL (McCormick & Cliff, 1977) is used for individual administration and the FORTRAN version (Cudeck, Cliff, & Kehoe, 1977) is designed for group administration with a minimum of 15 examinees. As more students are tested, more accurate difficulty estimates are made, thereby resulting in a more individually tailored test administration. McCormick and

Cliff (1977) claim that after six administrations, there is a significant reduction in the number of items administered to subsequent examinees.

The FORTRAN version of TAILOR begins by administering the same item to all examinees. An implied ordering of items is performed based on the observed numbers of correct and incorrect responses. Examinees' responses are also used to award examinees with a correct response to easier items. The process continues by matching item difficulty estimates with the examinee's performance on previous item.

The Broad Range Tailored Test of Verbal Ability (BRITTVA) (Lord, 1977) is an excellent example of an adaptive test administration system. While implementing adaptive testing strategies, the BRTTVA can be used to assess verbal ability from the fourth grade level to the graduate level. In addition, parallel test forms may be generated.

Nonadaptive testing. The first example is a system used by the School of Basic Medical Sciences at the University of Illinois, Urbana-Champaign. Students are directly involved in the administration of diagnostic assessment examinations. Students must take nine comprehensive exams each containing approximately 180 items covering an individual clinical problem. Four to five hours are needed to complete the exam. Since students work through the curriculum at their own pace, all students will not necessarily be taking the examination at the same time.

The test administration system, named LEVEL3, is written in TUTOR for PLATO IV implemented on a CDC 7600 (Sorlie, Essex, & Shatzer, 1979). Exams are administered via a PLATO IV terminal. First, students must schedule their exam using the "Level III Scheduler" program. Students specify a test date and time as well as total testing time required. At the scheduled time of the test, the student logs on the system and is presented with examination instructions. Once the exam is specified, a list of disciplines covered in the exam is presented to the users, which includes the number of items within each discipline. The student then specifies the sequence in which he/she would like the disciplines to be presented. During the exam, the student still has some control over the order of administration. Students have the option of omitting items and receiving a zero score, or skipping an item and returning to it after attempting all items in the current discipline. At the end of each discipline, students' scores are presented to them. Before proceeding with the next discipline, questions answered incorrectly may be reviewed and a second answer may be selected. This process allows students to raise their scores. The scoring component of this system keeps track of all student responses including a record of items answered correctly on a second attempt.

Most test administration applications for a microcomputer can be classified as either page-turners or drill and practice exercises. In this format, test items are presented on the monitor and examinees respond to items one at a time. Other

than that, features of traditional paper and pencil testing remain. A more appropriate use of microcomputers for administering tests is currently being field tested at several colleges across the country. Educational Testing Service (ETS) and the College Board have developed a system for administering both conventional and adaptive tests (Ballas, 1984). The current emphasis is to provide institutions with a tool for administering and scoring placement tests in a relatively short period of time.

There are a few recently developed test administration systems which use a somewhat traditional format in the administration of the test, while applying many of the advances in computer technology to the scoring and reporting aspects of the test. Although the development of these systems was not centered around new applications of measurement theory or new measurement theories, they do seem to illustrate the newest trend in computerizing test administration.

KEYWAY, a test scoring and reporting system developed by ETS, is characterized by many of the advantages of computer-assisted test administration systems, cited previously. Rather than employing computers in the test administration phase, however, KEYWAY uses microcomputers in recording answers, scoring, and reporting results. The Center for Occupational and Professional Assessment (COPA) uses KEYWAY for several licensing examinations, including the Real Estate Licensing Examination (RELE). When preregistered candidates report to a KEYWAY testing center, they receive a standard, printed test booklet

and a KEYWAY Answer Pad. The Answer Pad is not used to display the test content. Rather, its primary purpose is to record information. The one-line LCD panel displays each item number and waits for the students to respond with either an answer choice or a request to advance to the next item. Once all items have been attempted, the candidates may review all skipped items. At the end of the test, all demographic information and item responses are transferred to the Memory Module which is a portable, transferrable unit that resides in the Answer Pad during the exam. Upon completing the exam, the candidate returns the Answer Pad to the test administrator. Scoring is accomplished by removing the Memory Module from the Answer Pad and inserting it into the Memory Reader which reads the responses and downloads the information to an IBM PC. The test is then scored and a printout of results is produced.

During the fall of 1985, COPA pilot tested a second computer-assisted test recording, scoring, and reporting system which will be available for candidates of the National Association of Purchasing Management (NAPM) Examination. Although quite similar to KEYWAY, this system has some unique features. The hardware requirements for this sytem include the Radio Shack TRS 80 Model 2 or Model 12, a printer, and two integral or separate disk drives. Although candidates receive standard printed test booklets, directions and items also appear on the monitor. Rather than using a Memory Module type device, standard floppy disks are used for permanent storage of test information. When each module of the exam is completed, the

test is scored and two copies of the score report are produced - one for the candidate and one for ETS use.

For the past two years, the American College has been offering computer-based examinations through their Examinations on Demand (EOD) program. Prior to 1982, students enrolled in the Chartered Life Underwriter (CLU) or the Chartered Financial Consultant (ChFC) program, were required to pass 10 nationally administrated paper-and-pencil examinations. These exams were offered twice a year. Since the EOD program began, candidates have the option of taking the tests in standard written format on the predetermined test dates or requesting the computer-administered version of the exam at a time which fits their own schedules.

The EOD exams are administered through the Control Data Corporation (CDC) Education Center network, which houses PLATO terminals. Through the CDC network, candidates have substantially more flexibility in scheduling both the time and location of the exam, since Education Centers are currently located in 35 states across the country.

Similar to many computer-administered testing systems, the CLU and ChFC exams are scored immediately upon completion and the candidates leave the test site with their scores in hand. Nungester and Vaas (1984a) report that in the first two years of the program, over 15,000 candidates have participated in the EOD program. Characteristics of participants in the EOD program are continually being examined (Nungester & Vaas, 1984b). It is hoped that by examining the characteristics of students who opt

for the EOD system, insight may be gained as to the acceptance, advantages, and disadvantages of the current system.

It should be noted that the computerized version of The American College tests employs item selection strategies based on both item difficulty and test content.

IV. EVALUATION AND RESEARCH

The purpose of this section is to discuss issues or studies dealing with the evaluation of test construction systems as well as using the proposed testing systems to study test construction problems. If the studies investigate testing issues that may have implications for computer-assisted test construction, though they may not involve any computer-assisted test construction systems, they are included in this section.

To use computers to assist in test construction, we must make sure the quality of the test construction process will not be compromised. Therefore, evaluation of any test construction application is not only necessary, but also indispensable. Unfortunately, thorough evaluation and research are not usually done before a system is on the market for distribution (Deck & Estes, 1984). This could be attributed to several factors. First of all, to sell a product in a competitive market, timing is very crucial. Most developers cannot wait for a long delayed evaluation to be carried out. Secondly, the technology is changing so rapidly. If a system is not on the market right away, new technology may make the system obsolete. Another reason could be the users' fault. Users are so fascinated by the technology, they tend to ignore how the system contributes to educational testing.

A computer-assisted test construction system must be evaluated from three viewpoints. These three perspectives, in order of priority, are (a) the measurement specialists' view, (b) the users' (teachers') view, and (c) the computer

specialists' view. Measurement specialists must make sure the quality of the test construction process is maintained or improved. If a system is not theoretically sound from a measurement perspective, it should not be introduced to the users. Users are responsible for determining whether the system is appropriate for their intended clientele in terms of ease of use and meeting their needs. Computer specialists' responsibilities are to make sure the system is running smoothly and efficiently. But efficiency should not override measurement quality and usability.

As with any product, two different evaluations should be conducted: formative and summative. During the formative evaluation, data should be collected for the purpose of improving the system to make sure it functions as intended. The summative evaluation should consider whether the implementation of the system improves the quality of the test construction process by the users. Both types of evaluation data should be available to the clients before a product is distributed on a large scale.

So far, with the exception of adaptive testing, computer-assisted test construction systems' evaluation data are rather scarce. Although published reviews are available from some journals (e.g., Educational Technology, Social Science Microcomputer Review), they cannot be used as a substitute for a formal evaluation. Normally, these reviews are only intended to serve as a buyer's guide. To examine the quality of a system, users must demand evaluation data obtained through a formal

process. Appropriate evaluation instruments, such as the one cited in Hsu and Nitko (1983), or the user's evaluation form proposed by Ju (1984) should be used to collect evaluation data. Section I of the user's evaluation form proposed by Ju is given in Appendix A. The 30 statements were designed to measure the following aspects of the computer package: usefulness, efficiency, documentation error handling, and performance. Section II is a series of open-ended questions. This section allows users to add any additional comments wnich are not addressed in Section I.

Whether the capabilities of computers have been used efficiently and effectively can be evaluated by computer specialists. Consumers of testing systems must consider both hardware and software capabilities in terms of intended uses. Readers interested in criteria for hardware and software selections may wish to consult guidelines published in various journals (e.g., Hiscox, 1983, 1984a).

Item Construction

Since this area is still in a rather primitive stage, literature concerning the evaluation and research of computerized item writing procedures is relatively scarce. In this section, we will first briefly illustrate relevant evaluation issues by using the works of Millman (1982). Then the focus will turn to research on item writing techniques which may not yet be computerized.

The computer-based test construction system contructed by Millman and Outlaw (1978) was implemented in an introductory statistics course using a mastery learning strategy. The evaluation focused on both the system and student attitudes and learning. Millman (1980) reported that the computer programs had produced all features anticipated and then ran smoothly without any detectable errors. Major drawbacks of the system concern the specific configuration of computer hardware and poor documentation. These drawbacks limited the transportability of the system.

Using such a system produced some positive impacts both in instructional processes and student attitude. The instructor had to prepare more thoroughly in terms of what should be taught and assessed. The students showed positive attitudes toward the mastery test approach. Final examination scores of students involved in this approach, however, failed to demonstrate superiority in comparison with the scores of students involved in the traditional approach. The researcher has attributed this finding to the limitation of the criterion measure employed. Unfortunately, the quality of items generated by the system was not reported. The evaluation of the system would be even better if other instructors were involved.

Research issues concerning item writing techniques can be illustrated by the study conducted by Roid and Finn (1978). To assess the feasibility of employing the linguistic transformation approach in item construction, computer-based algorithms were developed and used to analyze prose subject

matter. High information words were identified from prose. Sentences containing these words were transformed into multiple-choice items by item writers who generated alternatives using an informal approach and by an algorithmic approach. Items from these two approaches were compared and evaluated using data obtained from the try-out of the items. Results showed that both types of items were equally effective in measuring learning. Items derived from key word nouns tended to produce low quality items. The authors concluded that the algorithmic approach is feasible in generating foils for multiple-choice items.

Before item writing techniques can be actually computerized, more studies like Roid and Finn (1978) are needed to determine which aspects of each item writing technique can best be done by a computer and which aspects the computer cannot perform as well. We should implement only the ones that can produce quality items.

Item Banking

Researchers interested in item bank evaluations may wish to check the following sources for ideas: Hiscox and Brzezinski (1980), Hiscox (1983), Deck and Estes (1984), Estes and Arter (1984), Baker (1972), Millman and Arter (1984), and Hsu and Nitko (1984). These studies do not deal directly with item bank evaluation. However, in their discussions of the requirements for a good item bank or in their reviews of item banks, they

mention criteria that may be useful in item bank evaluation. If we know the requirements for a good item bank, we should be able to identify the criteria that can be used for evaluation. Since some of the criteria are discussed in an earlier section, they are not repeated here.

How feasible is the use of an item bank for test development? This issue was investigated by Brzezinski and Demaline (1982). After comparing test development under the traditional approach with test development using an item bank in terms of both costs and outcomes of test development, they concluded that more advantages can be gained by using an item bank. We have to keep in mind that these two different approaches are not directly comparable.

Proper use of item banks depends on their intended use and the quality of items stored in the banks. Without good quality items, item bank applications are not likely to produce good outcomes. In addition to test assembly for individual instructors, item banks have been used to monitor pupil performance (e.g., Johnson and Maher, 1982), and to establish and maintain cutoff scores for teacher certification examinations (Legg, 1982). One application that is focused on here is the use of item banks in adaptive testing. The most common application is to use estimates obtained from item response theory as indicators of the quality of items used in the bank.

As shown by Jensema (1977), Bayesian decisions were affected by the characteristics of the item bank. What

characteristics of the item bank should be of primary concern? Most research has focused on the following main issues:

(1) What is the minimum number of items required for the bank?

(2) Should the one- or three-parameter model be used?

(3) What is the minimum sample size required to calibrate the items?

(4) What is the minimum number of items required for a test?

Since these four issues are related to each other, our discussions of these issues obviously cannot be completely separated one from the other.

The size of item banks varies greatly from one to another. Examples cited by Wright and Bell (1984) range from 51 items to 9452 items. Naturally one may wonder what minimum number of items is required for an item bank. The issue will not be an issue if good items are available. For the reason of economy, however, users of item banks may not be able to store as many items as desired. Also the retrieval process will be slowed down substantially and the classification procedure will be very complicated when a large item bank is involved. On the other hand, too few items are more likely to create serious problems for testing than too many items. Therefore, the general rule of "the more the better" (Millman and Arter, 1984) seems reasonable.

Most studies concerning this issue are usually in the context of adaptive testing. Ree (1981) conducted a simulation

study to investigate the effects of item calibration, sample size, and item pool size on adaptive testing. Using the three-parameter model, calibrated item pools of 100, 200, and 300 items with calibration sizes of 500, 1,000, and 2,000 were examined. Based on the reduction of absolute error of ability estimates, he concluded that a minimum of 200 items with calibration size of 2,000 subjects is required. Sizes between 200 items and 100 items may be adequate if the items have high discrimination power and a wide range of difficulty (Reckase, 1981). Urry (1977) also emphasized the quality of items in the bank. In addition to the requirement for a minimum of 100 items, the item discrimination parameter must exceed .8 and the item difficulty parameter must be spread evenly and widely.

Weiss and Kingsbury (1984) also concurred that a minimum of 100 items is acceptable. Green et al. (1982) indicated that the United States Armed Services are planning to develop a pool of 200 items for each of the ten proposed computerized adaptive Armed Services Vocational Aptitude Battery tests. Since unidimensionality is assumed for IRT, this item pool size should be considered as the minimum requirement for measuring one particular trait. The minimum number of items required for an item bank can be determined if the number of traits or contents to be measured is decided.

Item statistics are useful indicators of the quality of item. We should not ignore them simply because they may be misused by users who believe only in statistical criteria in judging item. Because of this position, our concern is the

issue of what kinds of item data should be collected rather than the issue of whether items should be calibrated as discussed by Millman and Arter (1984).

Three kinds of item data are most commonly used: classical item statistics, estimates based on the Rasch model (1PL) and estimates based on the three-parameter logistic model (3PL). One of the major factors to be considered in deciding which kind of statistics to be used is probably the sample size available for item calibration. For small classroom testing, calibration of items based on item response theory is not possible. Some classical item statistics may be appropriate for small classroom testing (Nitko & Hsu, 1983). These statistics may be computed and stored along with the items in item banks. For calibration using the 3PL model, 1,000 subjects per item is required (Green et al., 1982, 1984; Weiss & Kingsbury, 1984). However, Hambleton and Cook (1983) have shown that for a 20 item test, the increase in the precision of the standard error of ability estimate from a calibration size of 200 to 1,000 is relatively small. When the sample size is less than 200, Lord (1983) showed that the Rasch model performs slightly better than the two-parameter model.

It seems reasonable to say that the superiority of the 3PL model cannot be exhibited unless a large calibration sample is available. Also, if the item pool is small, say 40, the 3PL model does not show any advantage in tailored testing either (McKinley & Reckase, 1984). These results and other complications in using the 3PL model, such as the possibility of

non-convergence, may lead one to conclude that the 1PL model should be used instead of the 3PL model. Such a conclusion obviously is premature. These results only imply that the 3PL model is not any better than the 1PL model when they are compared under less than desirable conditions. Also, these results do not prove that the 1PL model is accurate. Instead, these results only imply that if the 3PL model is employed under desirable conditions, the results obtained have a better chance of being accurate.

Test Design

Test design using traditional item statistics does not really require computer assistance. Most users subjectively decide whether or not an item should be included after examining item information. At most, they may estimate the reliability coefficient of the newly designed test using item statistics obtained previously. This process has been computerized by some systems (e.g., Nitko & Hsu, 1984a). Nevertheless, we cannot find any research studies dealing specifically with this issue.

Computer assistance is most likely to be required if item response theory is utilized in test design. Since such an application is relatively new, there are rather few studies available in this area. The only example which we may discuss is the system developed by Sadock (1984), which was illustrated previously. In that system, the developer utilized four measurement specialists and nine users (teachers) to tryout the

system. In addition to comment on specific sections of the system, time required for test design was also recorded and analyzed. At the end, the users also completed a rating form to reflect their impressions of the system.

In addition, to fully assess the capacity and efficiency of the system, as well as the quality of the tests produced, the researcher designed 144 experimental tests. For these trial runs, the following test characteristics were varied: (a) type of test or intended use of test results (i.e., tests were designed either to group students, rank students, or assess individual student mastery); (b) degree of specificity of the content domain (which affects the size of the item pool); (c) desired test length; (d) the number of and value of the cutoff-score(s) when classification tests were developed; and (e) the range of class ability level when a test designed to rank students was specified. For each experimental test, length of time required to design the test was recorded. Test information curves were plotted and compared to the theoretical test information curve for each type of test, to determine whether maximum information was indeed obtained at the critical decision-making point(s) on the ability scale. Although the evaluation study was quite extensive in coverage, the system was not evaluated by computer specialists.

Test Administration

Research concerning computerized test administration may be classified into two major categories: (a) research designed to

compare computerized testing with conventional testing procedures, and (b) research designed to study adaptive testing strategies which may be computerized.

Roid (1984b) listed 11 studies published between 1969 through 1984 comparing computerized testing with conventional testing. Tests involved were standardized intelligence test and personality inventories. Most of the studies found no significant difference between the two testing modes. Some of the major findings include: (a) a high state of anxiety under computerized testing; (b) more honesty, openness, and willingness to respond under computer administration; and (c) the detection of unexpected responses under computerized testing. In general, subjects' familiarity with the computer seems to affect their performance under computerized testing.

One major advantage of computerized testing is the ability to adapt the test to subjects' ability level. Since studies listed in Roid (1984b) are nonadaptive, nonsignificant findings in most studies should not be a surprise. Merely simulating paper and pencil tests on computers is not a good way to utilize computer technology. Unless computerized testing can do a better job than regular paper and pencil testing, it is not justifiable to use expensive computer testing to replace relatively inexpensive paper and pencil testing which can be administered in a large group simultaneously.

Since so many studies on adaptive testing strategies have been generated during the last decade, it is not possible to cover them in a few pages. Instead, only a few representative

studies are discussed. The focus here is on strategies of item selection when an adaptive test administration approach is employed.

As mentioned previously, computerized adaptive testing was most successfully applied to aptitude testing. In this section, we discuss two studies dealing with the evaluation of such an application. The first is an empirical investigation of the Broad Range Tailored Test of Verbal Ability (BRTTVA) developed by Lord (1977), which was described in a previous section. This investigation was conducted by Kreitzberg and Jones (1980). To carry out the study, the researchers developed a computer system that can administer two forms of the BRTTVA. Each form of the test consisted of 25 items and the administration of the two forms was counterbalanced. The BRTTVA was administered to 146 high school students. A questionnaire to measure examinees' attitudes toward the testing was administered at the ends.

Data analyzed and presented included descriptive characteristics of the observed data, information functions of both forms, reliability and validity, and the performance of the maximum-likelihood estimators. Since this estimator is the key to the selection of items, a brief description of its performance is warranted. The item selection procedure was investigated by a Monte Carlo analysis. To compare the actual item selection procedure and the ideal situation, scattergrams were plotted to show the relationship between the number of correct responses and the final estimate of ability. Ideally, no regression would be expected. The results show some

regressions. In addition, these graphs were also compared with graphs obtained by simulating the responses of examinees using the estimated ability. Some discrepancies were noted. This implies that the item selection process is in need of further improvement.

The second example is an evaluation plan developed for the Navy by Green, et al. (1982). Suggested areas of evaluation include item content, reliability, validity, item parameters, item pool characteristics, item selection and test scoring, stopping rules and so on. In terms of item selection, three methods were suggested: the Bayes updating method proposed by Owen (1969, 1975), the maximum information method proposed by Lord (1977), and a finite Bayes method proposed by Bock and Aitkin (1981).

Although this report (Green et al., 1982) includes only an evaluation plan, its recommendations for evaluations should be considered seriously by anyone who is planning to develop an adaptive testing system. The recommendations cited below address the efficiency of item selection in adaptive testing:

1. "The prccedure for item selection and ability estimation must be documented explicitly and in detail." (p. 52)
2. "The procedure should include a method of varying the items selected, to avoid using a few items exclusively." (p. 52)
3. "The procedure used should include a mechanism to maintain a rough balance of correct answer options." (p. 52)

4. "The computer algorithm must be capable of administering designated items, and recording the response separately, without interfering with the adaptive process." (p. 53)

5. "The computer system must be able to base the choice of a first item on prior information." (p. 53)

The nature of research on adaptive testing strategies during the 1970s is probably best represented by the final report prepared by Weiss (1976). Since 1973, a group of researchers at the University of Minnesota, under the leadership of David Weiss, has dedicated itself to the study of computerized testing and produced many technical reports. This 1976 final report is a summary of their efforts during the first three years. The objectives of their investigation were: (a) to develop and implement the stratified adaptive ability testing using computers, (b) to compare various strategies for adaptive testing, (c) to study the effect of item selection and feedback on ability test scores, and (d) to assess the usefulness of test information for diagnostic purposes.

Included among the 21 major findings presented by Weiss (1976) are the following:

(a) The rankings of adaptive strategies, in terms of logical analysis, are Bayesian [illegible]num likelihood, stradaptive, pyramidal models, and [illegible]level.

(b) Based on information curves, stradaptive and Bayesian are most desirable and flexilevel is least desirable.

(c) The Bayesian approach has certain weaknesses that will limit its utility.

(d) In addition to its logical appeal, simulation results show that "the stradaptive test appears to provide the best realization of the ideal of measurement with equal and high precision of all trait levels" (p.3)

Since our interest in adaptive test administration is limited to item selection strategies, we will discuss only two additional research issues below. Readers interested in other issues of adaptive testing may wish to consult Clark (1976) and Weiss (1974, 1978, 1980, 1983).

The first issue to be addressed is related to the Bayesian decision strategy. This strategy has attracted a great deal of interest since the publication of Owen's studies (1969, 1975). Weiss (1974) and his associates (Vale & Weiss, 1975; McBride & Weiss, 1976) have found some strengths and some weaknesses. In terms of desirable characteristics of adaptive testing and information curves, the Bayesian strategy is ranked as one of the highest among the various strategies compared. However, the obtained ability estimates were found to be highly correlated with test length. Although the estimates are not equally precise throughout all ability levels, the obtained scores seem to be related to the prior ability estimate used.

This strategy has been implemented by Urry (1975). The results seem promising, but the strategy has never been evaluated under real life testing situations on a large scale. With the development of computer technology, the computational

aspects of the strategy should be feasible even using microcomputers. Further research on this strategy should be encouraged.

The second issue is related to adaptive achievement testing. A great deal of theoretical research has been done on adaptive aptitude testing. Relatively little research has been done on achievement testing, possibly due to the difficulty in dealing with multi-trait assessment in item response theory. Recent developments in multivariate methods should be able to provide a theoretical foundation for multi-content achievement testing (Roid, 1984b, Embretson, 1985). Without a theoretical base, earlier attempts to investigate adaptive achievement testing usually made decisions about each objective independently. Branching between objectives was established in advance either through logical analysis or hierarchical analysis. Weiss and his associates (Brown & Weiss, 1977; Gialluca & Weiss, 1979) proposed an inter-subject branching strategy for achievement testing. More theoretical studies in this are are badly needed. Since achievement testing is so important in the educational enterprise, adaptive achievement testing should have a great deal of potential. Item selection strategies for this type of achievement testing should prove to be a challenging issue.

To evaluate computer administered tests, a set of criteria proposed by Millman (1984) should be considered. These criteria are: (a) cost efficiency, (b) comparability to measure what is desired, (c) feasibility, (d) contribution to instruction,

(e) precision of measurement: testing time, (f) security, (g) concern for the individual, (h) ease of scoring, and (i) fairness. This set of criteria should be applicable to both adaptive and nonadaptive testings. In addition, guidelines being prepared by the American Psychological Association (1985) should be considered.

V. PROSPECTS FOR THE FUTURE

Many attempts have been made to explore the possibility of using computer technology to assist in the construction of tests during the last 20 years. Actual successful applications on a large scale are relatively few. For those applications that are actually in operation, most are simply replacing paper and pencil tests or human labor by the computer. With the exception of adaptive testing, it is very difficult to find any documentation which shows that the quality of assessment processes is improved as a result of using the computer. Is the quality of items improved because we can generate and construct items by using the computer? Is the quality of tests improved because the computer can be used to bank items and/or to design tests? Is the quality of testing improved because of the feasibility of computerized adaptive testing? We have some positive evidence for the last question. But computerized adaptive testing procedures have not been implemented on a large scale and are limited to aptitude testing. This limited success could be attributed to various reasons (Roid, 1984b). One of the reasons is probably related to computer technology. Before microcomputers, the accessibility of computers was a problem for most users. The development of new hardware and the portability of software between different hardware are also causes for the limited implementation. This difficulty may be reduced with the availability of microcomputers. Another reason which we believe is very crucial, is an overwhelming neglect of assessing

measurement quality in applying computer technology. Although there are many testing packages available on the market, good products are difficult to find. We are so fascinated by the technology that we want to do everything using the computer. But we must ask whether the computer can improve the quality of the activity or not. It is all right to be concerned about the technical quality such as beautiful color graphics and short response time, but we must also pay attention to the quality from a measurement perspective. If we can overcome these difficulties, the future of computer-assisted test construction should be very bright.

In the remaining sections, an attempt is made to outline some prospects for the future of computer-assisted test construction.

Item Construction

Using the computer to construct items is very useful, but not easy. Although the success is rather limited at this point, the potential is relatively great. To be successful, we may work from various directions:

(a) The first priority is to develop more item construction theories that can take advantage of artificial intelligence and the phrase recognizability of the computer. More specifically, Millman (1980) suggested that we have to develop a high level computer language specifically for item writing purposes and improve our

domain specification strategies to make them feasible for computer item generation. (See also Roid, 1984b.)

(b) Effort should be made to develop different item types which can take advantage of the computer capability (Hambleton, 1984; Johnson, 1983; Wood, 1984). Traditional item types are designed for paper and pencil tests. Routinely using the computer to construct and administer items of the traditional types is not a desirable approach for using the computer. We must take advantage of the computer's capabilities to improve our testing by developing new item types. For example, we may be able to use graphics to simulate test item conditions for problem solving. Some prototype items have been developed already by Hunt and his associates (Hunt, 1985) to measure spatial ability.

(c) Studies should be conducted to evaluate parallel tests generated by the computer (Millman, 1980). Interpretation guidelines for parents concerning test results obtained from different forms of a test should be considered also.

(d) There is a need to develop software which can construct reading comprehension tests based on textbooks. (Roid, 1984a)

Item Banking

For an item bank to be used widely, public agencies and textbook publishers must be involved. Development of item banks in the area of criterion-referenced achievement tests should be encouraged. Item banks developed and distributed with textbooks by publishers are becoming popular these days (e.g., PRISM developed by the Psychological Corporation (1982) and Academic Institutional Measurement System (AIMS) distributed by Charles E. Merrill Publishing Co.). But we believe good general purpose item banking systems should also have good potential. The meaning of general purpose implies that all users can store their own items. The item classification scheme should be general enough for most purposes.

There is a possibility that general data base programs will be used more often in developing item banking programs (Deck & Estes, 1984). Since more computer knowledge is required in using a general purpose data base, this prediction may be true for professional test developers rather than for common test users.

Another approach which may be considered is to develop item form banks. Instead of storing items, item forms (or other item generation techniques) may be stored. Items will be generated by the computer when a specific form is selected. This approach combines item generation with item banking. It should eliminate the concern about storage space for items.

Test Design

Current test design using IRT depends on mainframe computers to calibrate the items first. Then, the estimates of item parameters are transferred to microcomputers. With the development of the second generation of microcomputers, test design using IRT may not require mainframe computers any more. MicroCAT (Assessment Systems Corporation, 1985) has already incorporated such a capability into their system. If items can be calibrated and the test can be designed in one system, IRT may be used to develop tests by teachers who know very little about measurement theories.

Another factor that will affect future test design is the new development in the area of multivariate methods. These methods make the study of the quality of test items using multidimensions possible (Roid, 1984b).

Test Administration

The development of test administration using computers depends on the development of good item banks. In order to speed up the process of developing good banks, textbook publishers and public agencies may have to be involved in the development of testing systems. With the increasing availability and capability of microcomputers, powerful item selection strategies, such as Bayesian and maximum likelihood, may be implemented and used in test administration on a large

scale. This is feasible because of the availability of microcomputers and the development in the area of microcomputer networking.

Further developments of psychometric theories in the areas of achievement and diagnostic testing are needed (McArthur & Choppin, 1984). These theories are needed to guide the anticipated popularity of computerized achievement and diagnostic testing. Also, there is a need to develop item analysis procedures based on data obtained from individualized testing. It seems illogical to use data obtained from group administered testing to estimate item parameters which are going to be used for individualized testing.

Implications for Education Testing

The impact of the development of computer-assisted test construction on testing is most likely to be felt in the following directions:

(a) Practicing computerized adaptive and diagnostic testing in classrooms, both in aptitude and achievement areas.

(b) Applying IRT in test design by non-measurement specialists.

(c) Using new item types and/or new assessment procedures in classrooms.

(d) Evaluating and using items distributed by textbook publishers rather than teachers writing their own items written by teachers.

(e) Increasing popularity of computerized certification and licensing testings.

APPENDIX A
SAMPLE OF A USER'S EVALUATION FORM
(Adapted from Ju, 1984)

Direction: To what degree would you agree or disagree with the following statements

	Strongly Agree	Agree	Disagree	Strongly Disagree	Don't Know
1. The package is useful for classroom testing.	4	3	2	1	0
2. Using the package is fun.	4	3	2	1	0
3. Using the package is frightening.	4	3	2	1	0
4. The package is "user-sensitive" or "user friendly".	4	3	2	1	0
5. Instructions to run the package are ambiguous and difficult to follow.	4	3	2	1	0
6. Using the package is boring.	4	3	2	1	0
7. The package runs smoothly.	4	3	2	1	0
8. It is hard to get back to the menu if the user makes a mistake.	4	3	2	1	0
9. Most important classroom testing activities are contained in the package.	4	3	2	1	0
10. The user can modify the program to fit individual needs.	4	3	2	1	0
11. The package allows users to repeat programs as often as they want.	4	3	2	1	0
12. The package includes too many programs dealing with trivial classroom activities.	4	3	2	1	0
13. The package adapts well to the user's requirements.	4	3	2	1	0
14. Complex computer skills are required to run the package.	4	3	2	1	0
15. Minimum training is needed to use the package.	4	3	2	1	0

		Strongly Agree	Agree	Disagree	Strongly Disagree	Don't Know
16.	Documentation is confusing and inconsistent with the package.	4	3	2	1	0
17.	Adequate documents are provided for running the package.	4	3	2	1	0
18.	All possible responses are anticipated to make the package's operation predictable and reliable.	4	3	2	1	0
19.	Incorrect inputs are detected by the package.	4	3	2	1	0
20.	Input alternatives are flexible.	4	3	2	1	0
21.	Output alternatives are flexible.	4	3	2	1	0
22.	Screen display is clear and easy to read.	4	3	2	1	0
23.	Output summaries are difficult to interpret.	4	3	2	1	0
24.	The package has many uncorrected "bugs" which cause it to behave inconsistently or to "crash".	4	3	2	1	0
25.	Feedback is ineffective and inappropriate.	4	3	2	1	0
26.	Overall the response time (the time lag between your request and the response by the computer) is reasonable.	4	3	2	1	0
27.	Error messages are confusing.	4	3	2	1	0
28.	Adequate procedures are incorporated to prevent the user's errors.	4	3	2	1	0
29.	The package achieves its intent.	4	3	2	1	0
30.	This is an excellent package; recommend without hesitation.	4	3	2	1	0

REFERENCES

American Psychological Association. (1985). Guidelines for computer-tests and interpretation. Revised Draft. Washington, DC: Author.

Anastasio, E. J. Marcotte, D. M., & Fremer, J. (1969). Computer-assisted item writing II (Sentence Completion Items (Test Development Memorandum 69-1). Princeton, NJ: Educational Testing Service.

Assessment Systems Corporation. (1985). User's manual for the MicroCATtesting system. St. Paul, MN: Author.

Baker, F. B. (1971). Automation of test scoring, reporting, and analysis. In R. L. Thorndike (Ed.) Educational Measurement, 2nd ed. Washington, DC: American Council on Education.

Baker, F. B. (1972). A convention item banking and test construction system. Proceedings of the Fall Joint Computer Conference.

Baker, F. B. (1973). An interactive approach to test construction. Educational Technology, 13, 13-15. ERIC No. EJ075561.

Baker, F. B. (1984). Technology and testing: State of the art and trends for the future. Journal of Educational Measurement, 21, 399-406.

Baker, F. B. (in preparation). Computer technology and testing. To appear in R. L. Linn (Ed.), Educational Measurement, 3rd ed. Washington, DC: American Council on Education.

Ballas, M. S. (Ed.). (1984). Computerized test system now in field. ETS Developments, 30, 1, 3.

Bejar, I. I., Weiss, D. J., & Kingsbury, G. G. (1977). Calibration of an item pool for the adaptive measurement of achievement. (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Berk, R. A. (1978). The application of structural facet theory to achievement test construction. Educational Research Quarterly, 3, 67-72.

Bock, R. D., Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Bowers, J. J. (1984). Database management for item banking and test development: An application of dBASE II for the microcomputer. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Bowers, J. J. (1985). dBASE III and dBANK: Further work in item banking for test development. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Brightman, H. J., Freeman, J. L., & Lewis, D. (1984). Constructing and using computer-based formative tests. Educational Technology, 24(6), 36-38.

Brown, W. A. (1973). A computer examination compositor for the IBM 360/40. Educational Technology, 13(3), 15-16. ERIC No. EJ075562.

Brown, J. M., & Weiss, D. J. (1977). An adaptive testing strategy for achievement test batteries. Research Report 77-6. Minneapolis: University of Minnesota, Department of Biology, Psychometric Methods Program. ERIC No. ED150165.

Brzezinski, E., & Demaline, R. (1982). An empirical investigation of the results of two different test development strategies. Portland, OR: Northwest Regional Educational Laboratory.

Byrne, C. (1976). Computerized question-banking systems: 1 - The state of the art. British Journal of Educational Technology, 7(2), 44-64.

Charles E. Merrill Publishing Co. (no date). AIMS: Academic Instructional Measurement System. Columbus, OH: Author.

Clark, C. L. (Ed.). (1976). Proceedings of the First Conference on Computerized Adaptive Testing. Washington, DC: U. S. Civil Service Commission. ERIC No. ED126110.

Cliff, N. (1973). Complete orders from incomplete data: Interactive ordering and tailored testing. Psychological Bulletin, 82, 289-302.

Computer Assisted Test Construction, special section (1973). Educational Technology, 13(3), 10-44.

Cudeck, R. A., Cliff, N., & Kehoe, J. F. (1977). TAILOR: A FORTRAN procedure for interactive tailored testing. Educational and Psychological Measurement, 37, 767-769. ERIC No. EJ174792.

Deck, D., & Estes, G. (1984). Microcomputer software support for item banking. Paper presented at the annual meeting of the American Educational Research Associaton, New Orleans.

DeGruijter, D. N. M. (1985). Item banking with dBASE II. Memorandum 842-85, unpublished report. Educational Research Center, University of Leyden, Leyden, The Netherlands.

Denney, C. (1973). There is more to a test pool than data collection. Educational Technology, 13(3), 19-20. ERIC No. EJ075565.

Embretson, S. E. (Ed.). (1985). Test Design: Developments in psychology and psychometrics. NY: Academic Press, Inc.

Estes, G. D., & Arter, J. A. (1984). Item banking for state and local test development and use. Portland, OR: Assessment and Evaluation Program, Northwest Regional Educational Laboratory.

Feldker, P. F. (1973). Computer-generated physics tests. Physics Teacher, 11, 304-305.

Ferguson, R. L., & Hsu, T. C. (1971). The application of item generators for individualizing mathematics testing and instruction. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Finn, P. J. (1975). A question writing algorithm. Journal of Reading Behavior, 7, 341-367.

Fremer, J. J., & Anastasio, E. J. (1969). Computer-assisted item writing: I, Spelling items. Journal of Educational Measurement, 6, 69-74.

Gialluca, K. A., & Weiss, D. J. (1979). Efficiency of an adaptive inter-subtect branching strategy in the measurement of classroom achievement (Research Report 79-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1982). Evaluation plan for the computerized adaptive vocational aptitude battery (Research Report 82-1). Baltimore: The Johns Hopkins University, Department of Psychology.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.

Haksar, L. (1983). Design and usage of an item bank. Programmed Learning and Educational Technology, 20, 253-262.

Hambleton, R. K. (Ed.). (1983). Applications of item response theory. Vancouver: Educational Research Institute of British Columbia.

Hambleton, R. K. (1984). Using microcomputers to develop tests. Educational Measurement: Issues and Practice, 3, 10-14.

Hambleton, R. K., Anderson, G. E., & Murray, L. N. (1983). Applications of microcomputers to classroom testing. In W. Hathaway (Ed.), New directions for testing and measurement: Testing in the schools. San Francisco: Jossey-Bass.

Hambleton, R. R., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing. NY: Academic Press.

Hazlett, C. B. (1973). MEDSIRCH: Multiple choice test items. Educational Technology, 13, 24-26. ERIC No. EJ075568.

Hiscox, M. D. (1983). A balance sheet for educational item bankers. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Hiscox, M. D. (1984a). A planning guide for microcomputers in educational measurement. Educational Measurement: Issues and Practice, 3, 28-34.

Hiscox, M. D. (1984b). Low cost educational test construction using microcomputers. Paper presented at Education Commission of the State Large-Scale Assessment Conference, Boulder.

Hiscox, M. D., Brzezinski, E. J. (1980). A guide to item banking in education. Portland, OR: Northwest Regional Educational Laboratory.

Hively, W. I., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 5, 275-290.

Holmes, S. E. (1983). CSAR: An interactive item bank system for the storage and retrieval of item information. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal. ERIC No. ED235233.

Howze, G. (1978). An interactive software system for Computer-assisted testing. AEDS Journal, 11, 31-37. ERIC No. EJ173608.

Hsu, T. C. (1975). Approaches to the construction of achievement test items. The Researcher, 14, 31-50.

Hsu, T. C., & Carlson, M. (1973). Test construction aspects of the computer assisted testing model. Educational Technology, 13(3), 26-27. ERIC No. EJ075569.

Hsu, T. C., & Nitko, A. J. (1983). Microcomputer testing software teachers can use. Educational Measurement: Issues and Practice, 2, 15-18, 23-30.

Hsu, T. C., & Nitko, A. J. (1984). A microcomputer item bank for use in classroom testing. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Hunt, E. (1985). Cognitive research that will change test design for the future. Paper presented at ETS Invitational Conference, New York.

Jensema, C. J. (1977). Bayesian Tailored testing and the influence of item bank characteristics. Applied Psychological Measurement , 1, 111-120. ERIC No. EJ 161401.

Johnson, J. W. (1983). Things we can measure through technology that we could not measure before. In R. Ekstrom (Ed.), Measurement, technology, and individuality in education. San Francisco: Jossey-Bass.

Johnson, S., & Maher, B. (1982). Monitoring science performance using a computerized question banking system. British Journal of Educational Technology, 13, 97-106. ERIC No. EJ268605.

Johnson, S., & Maher, B. (1984). A thesaurus-linked science question-banking system. British Journal of Educational Technology, 15(1), 14-23.

Ju, T. P. (1984). A formative evaluation of Pitt Educational Testing Aids by Teachers. Unpublished Master's Thesis, University of Pittsburgh.

Koch, W. R., & Reckase, M. D. (1978). A live tailored testing comparison study of the one- and three-parameter logistic models. (Research Report 78-1). Columbia, MO: University of Missouri, Department of Educational Psychology.

Koch, W. R., & Reckase, M. D. (1979). Problems in application of latent trait models to tailored testing. (Research Report 79-1). Columbia, MO: University of Missouri, Department of Educational Psychology.

Kreitzberg, C. B., & Jones, D. H. (1980). An empirical study of the Broad Range Tailored Test of Verbal Ability. RR-80-5. Princeton, NJ: Educational Testing Service.

Legg, S. M. (1982). The use of precalibrated item bank to establish and maintain cutoff scores. A case study of the Florida Teacher Certification Examination. Paper presented at the annual meeting of the National Council on Measurement in Education, New York. Eric No. ED221570.

Libaw, F. B. (1973). Constructing tests with the MENTREX tutorial testing system. Educational Technology, 13(3), 30-31.

Lippey, G. E. (Ed.). (1974). Computer-assisted test construction. Englewood Cliffs, NJ: Education Technology Publications. ERIC No. ED096948.

Lord, F. M. (1971). Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement, 31, 3-31.

Lord, F. M. (1977). A Broad-Range Tailored Test of Verbal Ability. Applied Psychological Measurement, 1, 95-100.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1983). Small N justifies Rasch model. In D. J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.

McArthur, D. L., & Choppin, B. H. (1984). Computerized diagnostic testing. Journal of Educational Measurement, 21, 391-397.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), New horizons in testing. New York: Academic Press.

McBride, J. R., & Weiss, D. J. (1976). Some properties of a Bayesian adaptive ability testing strategy (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

McCormick, D. J., & Cliff, N. (1977). TAILOR-APL: An interactive computer program for individual tailored testing. Educational and Psychological Measurement, 37, 771-774. ERIC No. EJ174793.

McKinley, F. L., & Reckase, M. D. (1984). An evaluation of one- and three-parameter logistic tailored testing procedures for use with small item pools. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Mediax Associates. (1984). MICROSCALE. Westport, CT: Mediax Associates.

Menne, J. W. (1973). Computer assisted test assembly at Iowa State University. Educational Technology, 13(3), 31-32. ERIC No. EJ075570.

Millman, J. (1974). Criterion-referenced measurement. To W. J. Popham (Ed.), Evaluating in education: Current applications. Berkeley: McCutchan.

Millman, J. (1980). Computer-based item generation. In R. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press.

Millman, J. (1982). A system for generating unique tests by computer and its use in a mastery learning setting. In P. R. Baumann (Ed.), Computer assisted instruction. Albany: State University of New York. Faculty grants for the improvement of undergraduate instruction Publicatoin series, Vol. 1.

Millman, J. (1984). Individualizing test construction and administration by computer. In R. Berk (Ed.), A guide to criterion-referenced test construction. Baltimore: The Johns Hopkirs University Press.

Millman, J., & Arter, J. A. (1984). Issues in item banking. Journal of Educational Measurement, 21, 315-330.

Millman, J., & Outlaw, W. S. (1978). Testing by computer. AEDS Journal, 11, 57-72. ERIC No. EJ177688.

Millman, J., & Westman. (personal communication)

Mizokawa, D. T., & Hamlin, M. D. (1984). Guidelines for computer managed testing. Educational Technology, 24(12), 12-17.

Nitko, A. J. (1983). Educational tests and measurement: An introduction. New York: Harcourt Brace Jovanovich, Inc.

Nitko, A. J., & Hsu, T. C. (1983). Item analysis appropriate for domain-referenced classroom testing (Technical Report No. 1). Pittsburgh, PA: University of Pittsburgh, Department of Educational Research Methodology.

Nitko, A. J., & Hsu, T. C. (1984a). Pitt Educational Testing Aids: User's Manual. (Development Edition). Pittsburgh, PA: University of Pittsburgh, Department of Educational Research Methodology.

Nitko, A. J., & Hsu, T. C. (1984b). A comprehensive microcomputer system for classroom teaching. Journal of Educational Measurement, 21, 377-390.

Nungester, R. J., & Vaas. C. E. (1984a). Nationally administered computer-based certification examinations: Overall design. Paper presented at the Fifteenth Annual Convention of the Northeastern Educational Research Association, Grossinger, New York.

Nungester, R. J., & Vaas, C. E. (1984b). Nationally administered computer-based certification examinations: Participant model. Paper presented at the Fifteenth annual convention of the Northeastern Educational Research Association, Grossinger, New York.

Oosterhof, A. C., & Salisbury, D. F. (1985). Some measurement and instruction related considerations regarding computer-assisted testing. Educational Measurement: Issues and Practice, 4(1), 19-23.

Owen, R. J. (1969). A Bayesian approach to tailored testing. Research Bulletin 69-92. Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the contect of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

The Psychological Corporation and Learning Achievement Corporation. (1982). PRISM. New York: The Psychological Corporation.

Reckase, M. D. (1981). Tailored testing, measurement problems and latent trait theory. Paper presented at the annual meeting of NCME, Los Angeles.

Reckase, M. D., & McKinley, R. L. (1984). Item difficulty reconsidered: An IRT perspective. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Ree, M. J. (1981). The effects of item calibration sample size and item pool size on adaptive testing. Applied Psychological Measurement, 5. 11-19.

Roid, G. H. (1984a). Generating the test items. In R. Berk (Ed.), A guide to criterion-referenced test construction. Baltimore: The Johns Hopkins University Press.

Roid, G. H. (1984b). Computer technology in testing. In B. S. Blake & J. E. Witts (Eds.), The future of testing: The second Buros-Nebraski symposium on measurement and testing. Hillsdale, NJ: Earlbaum.

Roid, G. H., & Finn, P. J. (1978). Algorithms for developing test questions from sentences in instructional materials. NPRDC TR 78-23.

Roid, G. H., & Haladyna, T. M. (1982). A technology for test-item writing. NY: Academic Press.

Sadock, S. F. (1984). A microcomputer-assisted test design system using item response theory. Doctoral dissertation, University of Pittsburgh.

Samejima, F. (1977). A use of the information function in tailored testing. Applied Psychological Measurement, 1, 233-247.

Sampson, J. P., Jr. (1983). Computer-assisted testing and assessment: Current status and implications for the future. Measurement and Evaluation in Guidance, 15(4), 293-299.

Seely, O. Jr., & Willis, V. (1976). SOCRATES' test retrieval at the California State University and Colleges. ASEA Journal, 9, 65-70. ERIC No. EJ139348.

Silvertson, S. E., Hansen, R. H., & Schoenenberger, A. O. (1973). Computerized test bank for clinical medicine. Educational Technology, 19, 29-34. ERIC No. EJ214671.

Sorlie, W. E., Essex, D., & Shatzer, J. (1979) Computer automated from sign-on to item analysis: A student appraisal system. Educational Technology, 19, 29-34. ERIC No. EJ214671.

Stocker, W., Esterson, C., & Schmid, R. (1977). A conversational random access computer-assisted test construction system. AEDS Monitor, 15, 6-7, 15. ERIC No. EJ159162.

Stocking, M. L., & Swanson, L. (1979). Computerized adaptive testing. CATC Digest, 3, 3-4.

Thissen, F. (1976). Information in wrong responses to the Raven Progressive Matrices. Journal of Educational Measurement, 13, 201-204.

Toggenburger, F. (1973). Classroom Teacher Support System. Educational Technology, 13(3), 42-43. ERIC No. EJ075577.

Urry, V. W. (1970). A Monte Carlo investigation of logistic mental test models. Doctoral dissertation, Purdue University.

Urry, V. W. (1975). Five years of research: Is computer-assisted testing feasible? Proceedings of the First Conference on Computerized Adaptive Testing. Professional Series 7556. Personnel Research and Development Center, U. S. Civil Service Commission.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 14, 181-196.

Vale, C. D. (1979). CIBEDS: Minnesota's item-banking system. CATC Digest, 3, 1.

Vale, C. D., & Weiss, D. J. (1975). A simulation study of stradaptive ability testing (Research report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. ERIC No. ED120209.

Vickers, F. D. (1979). Creative test generators. Educational Technology, 13,(3), 444. ERIC EJ05578.

Wald, A. (1947). Sequential analysis. New York. John Wiley & Sons.

Ward, W. C. (1984). Using microcomputers to administer tests. Educational Measurement: Issues and Practice, 3, 16-20.

Ward, W. H. (1984). BANKER: A test item storage, maintenance, and retrieval system for microcomputers. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Warm, T. A. (1978). A primer of item response theory. NTIS # AD-A063-072. Department of commerce, U. S. Coast Guard Institute, Oklahoma City, Oklahoma.

Wedman, J. F., & Stefanich, G. P. (1984). Guidelines for computer-based testing of student learning of concepts, principles, and procedures. Educational Technology, 24(12), 12-17.

Weiss, D. J. (1974). Strategies of adaptive measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (1976). Computerized ability testing, 1972-1975, Final Report. Washington, DC: Office of Naval Research. ERIC No. ED126131.

Weiss, D. J. (1978). Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (1980). Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (Ed.). (1983). New horizons in testing: Latent trait theory and computerized adaptive testing. NY: Academic Press.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21, 361-375.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST: User's guide. Princeton, NJ: Educational Testing Service.

Wood, S. (1984). Computer use in testing and assessment. Journal of Counseling and Development, 63, 177-179.

Woods, G. T. (1983). Computer-aided item development: An application of Samejima's new family of models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver: Educational Research Institute of British Columbia.

Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, how. Journal of Educational Measurement, 21, 331-345.

Wright, B. D., Mead, R. J., & Bell, S. R. (1980). BICAL: Calibrating items with the Rasch model. Research Memorandum 25C. Chicago: University of Chicago, Department of Education, Statistical Laboratory.

ERIC/TME Report 88

COMPUTER-ASSISTED TEST CONSTRUCTION: A STATE OF THE ART

By
Tse-chi Hsu
University of Pittsburgh
and
Shula F. Sadock
Pittsburgh Board of Education

As in other areas of education computer technology is being enthusiastically applied to many aspects of testing. While the promise and potenial of these applications to improve educational testing are great, they are not guaranteed.

This report reviews the literature on the utilization of computer technology to construct test items and/or to formulate tests according to sound measurement principles. It focuses on studies dealing with item construction, item banking, test design and test administration. Theoretical concepts, mainframe and microcomputer applications, evaluation and research issues and future prospects for computer assisted test construction are discussed.

RECENT TITLES
IN THE ERIC/TME REPORT SERIES

#91 - Evaluation of Corporate Training Programs, by Dale Brandenburg and Martin E. Smith. 5/86, $9.00.

#90 - Assessing Higher Order Thinking Skills, by C. Phillip Kearney, and Others. 4/86, $7.50.

#89 - Microcomputer Programs for Educational Statistics: A Review of Popular Programs, by Paul M. Stemmer and Carl F. Berger. 3/86, $7.50.

#88 - Computer-Assisted Test Construction. A State of the Art, by Tse-chil Hsu and Shula F. Sadock. 12/85, $7.50.

#87 - The Statewide Assessment of Writing, by Peter Afflerbach. 8/85, $7.50.

#86 - The Effects of Testing on Teaching and Curriculum in a Large Urban School District, by Floraline Stevens. 12/84, $6.00.

#85 - Reporting Test Scores to Different Audiences, by Joy A. Frechtling and N. James Myerberg. 12/83, $7.00.

#84 - Assessment of Learning Disabilities, by Lorrie A. Shepard. 12/82, $6.50.

#83 - Statistical Methodology in Meta-Analysis, by Larry V. Hedges. 12/82, $7.00.

#82 - Microcomputers in Educational Research, by Craig W. Johnson. 12/82, $8.50.

#81 - A Bibliography to Accompany the Joint Committee's Standards on Educational Evaluation, compiled by Barbara M. Wildemuth. 12/81, $8.50.

#80 - The Evaluation of College Remedial Programs, by Jeffrey K. Smith and others. 12/81, $8.50.

#79 - An Introduction to Rasch's Measurement Model, by Jan-Eric Gustafsson. 12/81, $5.50.

#78 - How Attitudes Are Measured: A Review of Investigations of Professional, Peer, and Parent Attitudes toward the Handicapped, by Marcia D. Horne. 12/80, $5.50.

#77 - The Reviewing Processes in Social Science Publications: A Review of Research, by Susan E. Hensley and Carnot E. Nelson. 12/80, $4.00.

#76 - Intelligence Testing, Education, and Chicanos: An Essay in Social Inequality, by Adalberto Aguirre Jr. 12/80, $5.50.

#74 - Intelligence, Intelligence Testing and School Practices, by Richard DeLisi. 12/80, $4.50.